Direct Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices

Olivier Ledoit[1] and Michael Wolf[1]

[1]Department of Economics
University of Zurich

Outline

Outline

The Problem

Problem:

- We want to estimate a p-dimensional covariance matrix based on an i.i.d. sample of size n
- The classic estimator is the sample covariance matrix S_n
- However, this estimator is ill-conditioned when p is of the same magnitude as n, and tends to perform poorly

The Problem

Problem:

- We want to estimate a p-dimensional covariance matrix based on an i.i.d. sample of size n
- The classic estimator is the sample covariance matrix S_n
- However, this estimator is ill-conditioned when p is of the same magnitude as n, and tends to perform poorly

Note:

- One of the most important problems in multivariate statistics
- Applications are plentiful

Previous Approaches

(1) Incorporate additional knowledge in the estimation process:

- Rely on a sparsity, such as Bickel and Levina (2008, AoS)
- Rely on a graph model, such as Rajaratnam et al. (2008, AoS)
- Rely on a factor structure, such as Fan et al. (2013, JRSS-B)

Previous Approaches

(1) Incorporate additional knowledge in the estimation process:

- Rely on a sparsity, such as Bickel and Levina (2008, AoS)
- Rely on a graph model, such as Rajaratnam et al. (2008, AoS)
- Rely on a factor structure, such as Fan et al. (2013, JRSS-B)

(2) Linear shrinkage:

- Consider estimators of the form:

$$\delta \cdot \bar{s}_n^2 \cdot I_p + (1 - \delta) \cdot S_n$$

where $\bar{s}_n^2$ is the grand mean of the sample variances $s_{n,i}^2$

- Ledoit and Wolf (2004, JMVA) derive asymptotically optimal *bona fide* shrinkage intensity δ under the Frobenius loss

Linear Shrinkage

Immediate interpretation:

- Shrink the elements of S_n to the elements of $\bar{s}_n^2 \cdot I_p$ with common intensity δ

Linear Shrinkage

Immediate interpretation:

- Shrink the elements of S_n to the elements of $\bar{s}_n^2 \cdot I_p$ with common intensity δ

Alternative interpretation:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$ with common intensity δ:

$$\lambda_{n,i}^{\text{shrunk}} := \delta \bar{\lambda}_n + (1 - \delta) \lambda_{n,i}$$

Linear Shrinkage

Immediate interpretation:

- Shrink the elements of S_n to the elements of $\bar{s}_n^2 \cdot I_p$ with common intensity δ

Alternative interpretation:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$ with common intensity δ:

$$\lambda_{n,i}^{\text{shrunk}} := \delta \bar{\lambda}_n + (1 - \delta) \lambda_{n,i}$$

In particular:

- The shrunken eigenvalues are obtained by applying a linear transformation to the sample eigenvalues

Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$, but at *distinct* intensities (even allowed to be negative)

Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \dots, \lambda_{n,p}); (u_{n,1}, \dots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$, but at *distinct* intensities (even allowed to be negative)

In particular:

- The shrunken eigenvalues are obtained by applying a nonlinear transformation to the sample eigenvalues

Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$, but at *distinct* intensities (even allowed to be negative)

In particular:

- The shrunken eigenvalues are obtained by applying a nonlinear transformation to the sample eigenvalues

Doing so should yield even better results, if done right.

Outline

Reasonable Restriction

Rotation-Equivariant Estimators

- Y_n are the observed data, an $n \times p$ matrix
- W is a $p \times p$ orthogonal matrix
- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is an estimator of Σ_n
- It is rotation-equivariant if $\widehat{\Sigma}_n(Y_n W) = W' \widehat{\Sigma}_n(Y_n) W$

Without specific knowledge about Σ_n, it is reasonable to restrict attention to this class of estimators.

Reasonable Restriction

Rotation-Equivariant Estimators

- Y_n are the observed data, an $n \times p$ matrix
- W is a $p \times p$ orthogonal matrix
- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is an estimator of Σ_n
- It is rotation-equivariant if $\widehat{\Sigma}_n(Y_n W) = W' \widehat{\Sigma}_n(Y_n) W$

Without specific knowledge about Σ_n, it is reasonable to restrict attention to this class of estimators.

We use the following class of rotation-equivariant estimators going back to Stein (1975, 1986):

$$\widehat{\Sigma}_n := U_n \widehat{\Delta}_n U_n' \quad \text{where} \quad \widehat{\Delta}_n := \mathsf{Diag}(\widehat{\delta}_{n,1}, \ldots, \widehat{\delta}_{n,p}) \text{ is diagonal}$$

Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to Σ_n
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1} \Sigma_n \widehat{\Sigma}_n^{-1}\right) / p}{\left[\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1}\right) / p\right]^2} - \frac{1}{\operatorname{Tr}\left(\Sigma_n^{-1}\right) / p}$$

Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to Σ_n
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1}\Sigma_n\widehat{\Sigma}_n^{-1}\right)\Big/p}{\left[\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1}\right)\Big/p\right]^2} - \frac{1}{\operatorname{Tr}\left(\Sigma_n^{-1}\right)/p}$$

Minimization problem :

$$\min_{\widehat{\Delta}_n} \mathcal{L}_n^{\mathrm{MV}}\left(U_n\widehat{\Delta}_nU_n', \Sigma_n\right)$$

Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to Σ_n
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1}\Sigma_n\widehat{\Sigma}_n^{-1}\right)\Big/p}{\left[\operatorname{Tr}\left(\widehat{\Sigma}_n^{-1}\right)\Big/p\right]^2} - \frac{1}{\operatorname{Tr}\left(\Sigma_n^{-1}\right)/p}$$

Minimization problem :

$$\min_{\widehat{\Delta}_n} \mathcal{L}_n^{\mathrm{MV}}\left(U_n\widehat{\Delta}_nU_n', \Sigma_n\right)$$

Solution:

$$\Delta_n^* := \operatorname{Diag}(\delta_{n,1}^*, \ldots, \delta_{n,p}^*) \quad \text{where} \quad \delta_{n,i}^* := u_{n,i}'\Sigma_n u_{n,i}$$

Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to Σ_n
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{\mathrm{Tr}\left(\widehat{\Sigma}_n^{-1}\Sigma_n\widehat{\Sigma}_n^{-1}\right)\Big/p}{\left[\mathrm{Tr}\left(\widehat{\Sigma}_n^{-1}\right)\Big/p\right]^2} - \frac{1}{\mathrm{Tr}\left(\Sigma_n^{-1}\right)/p}$$

Minimization problem :

$$\min_{\widehat{\Delta}_n} \mathcal{L}_n^{\mathrm{MV}}\left(U_n\widehat{\Delta}_nU_n', \Sigma_n\right)$$

Solution:

$$\Delta_n^* := \mathrm{Diag}(\delta_{n,1}^*, \dots, \delta_{n,p}^*) \quad \text{where} \quad \delta_{n,i}^* := u_{n,i}'\Sigma_n u_{n,i}$$

Note: Using the Frobenius loss instead yields the same solution.

Outline

Asymptotic Framework

Let $p := p(n)$ and assume $p/n \to c \in (0,1)$, as $n \to \infty$.

The following set of assumptions is maintained throughout:

- **A1** The population covariance matrix Σ_n is a nonrandom p-dimensional positive definite matrix.
- **A2** Let X_n be an $n \times p$ matrix of real i.i.d. random variables with zero mean, unit variance, and finite 16th moment. One observes $Y_n := X_n \Sigma_n^{1/2}$.
- **A3** Let $\{(\tau_{n,1}, \ldots, \tau_{n,p}); (v_{n,1}, \ldots, v_{n,p})\}$ denote the eigenvalues and eigenvectors of Σ_n. The e.d.f. of the population eigenvalues, denoted by H_n, converges weakly to some limit e.d.f. H.
- **A4** $\mathsf{Supp}(H)$, the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity. Furthermore, there exists a compact interval in $(0, +\infty)$ which contains $\mathsf{Supp}(H_n)$ for all large enough n.

Note: The paper also discusses an extension to the case $p > 1$.

Random Matrix Theory

A foundational result going back to Marčenko and Pastur (1967) states that the limiting distribution of the sample eigenvalues is deterministic

Under the stated assumptions, there exists a continuous limiting sample spectral distribution F such that $\forall x \in \mathbb{R}\;\; F_n(x) \xrightarrow{\text{a.s.}} F(x)$.

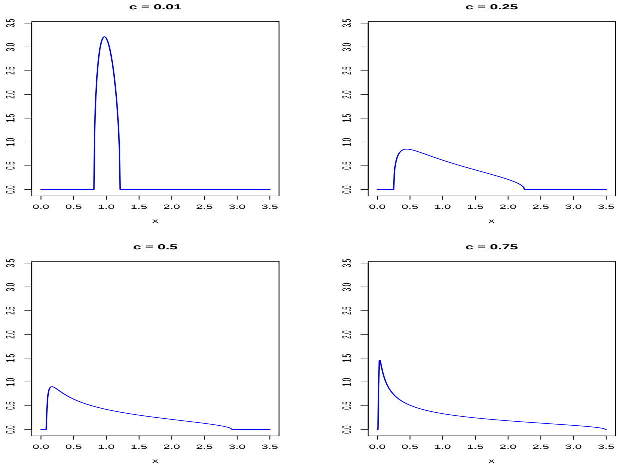
The limiting sample spectral c.d.f. F is uniquely determined by c and H; thus, we will refer to it as $F_{c,H} := F$ whenever clarification is needed.

A further implication is that the support of F is the union of a finite number of compact intervals.

Illustration

H is a point mass at one (such as for the identity covariance matrix).

Plot the density of F for various values of c:



Hilbert Transform

The Hilbert transform of a real function g is defined as

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_g(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x}$$

where PV denotes the Cauchy Principal Value.

It is thus the convolution of g with the Cauchy kernel $\frac{dt}{\pi t}$.
(Since the Cauchy kernel is singular, the integral does not converge in the usual sense and recourse to the Cauchy Principal Value is needed.)

Hilbert Transform

The Hilbert transform of a real function g is defined as

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_g(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x}$$

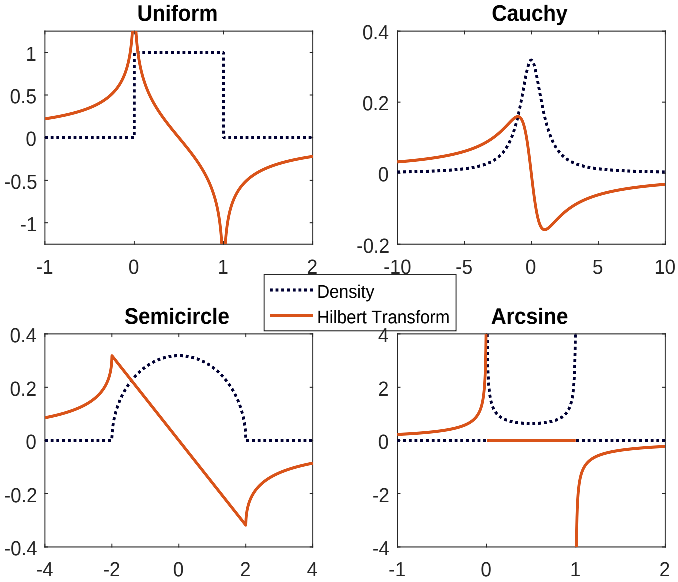
where PV denotes the Cauchy Principal Value.

It is thus the convolution of g with the Cauchy kernel $\frac{dt}{\pi t}$.
(Since the Cauchy kernel is singular, the integral does not converge in the usual sense and recourse to the Cauchy Principal Value is needed.)

Intuition:

- The Hilbert transform operates like a local attraction force
- It pushes x towards local mass centers
- For an illustration, plot the Hilbert transform of four densities

Illustration



Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of $\widehat{\delta}_{n,i}$ as $\widehat{\delta}_n(\lambda_{n,i})$, where $\widehat{\delta}_n$ is an unrestricted nonlinear shrinkage function, assumed to converge to a limit $\widehat{\delta}$.

Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of $\widehat{\delta}_{n,i}$ as $\widehat{\delta}_n(\lambda_{n,i})$, where $\widehat{\delta}_n$ is an unrestricted nonlinear shrinkage function, assumed to converge to a limit $\widehat{\delta}$.

Under the stated assumptions:

- For any $\widehat{\delta}$, the limiting loss of the estimator $\widehat{\Sigma}_n$ is deterministic
- One can minimize this deterministic limiting loss wrt $\widehat{\delta}$
- The solution yields an oracle nonlinear shrinkage formula

Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of $\widehat{\delta}_{n,i}$ as $\widehat{\delta}_n(\lambda_{n,i})$, where $\widehat{\delta}_n$ is an unrestricted nonlinear shrinkage function, assumed to converge to a limit $\widehat{\delta}$.

Under the stated assumptions:

- For any $\widehat{\delta}$, the limiting loss of the estimator $\widehat{\Sigma}_n$ is deterministic
- One can minimize this deterministic limiting loss wrt $\widehat{\delta}$
- The solution yields an oracle nonlinear shrinkage formula

The oracle formula is given by

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\text{o}}(x) := \frac{x}{\left[\pi c x f(x)\right]^2 + \left[1 - c - \pi c x \mathcal{H}_f(x)\right]^2}$$

where f denotes the density of F.

Nonlinear Shrinkage as Local Attraction

The oracle formula results in local attraction:
any sample eigenvalue is moved towards local mass centers.

Nonlinear Shrinkage as Local Attraction

The oracle formula results in local attraction:
any sample eigenvalue is moved towards local mass centers.

This phenomenon is easier to see based on the 'scaled' density $\varphi(x) := \pi x f(x)$, which yields the equivalent oracle formula

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\mathrm{o}}(x) = \frac{x}{1 + c^2\left[\varphi^2(x) + H_\varphi^2(x)\right] - 2cH_\varphi(x)}$$

Nonlinear Shrinkage as Local Attraction

The oracle formula results in local attraction:
any sample eigenvalue is moved towards local mass centers.

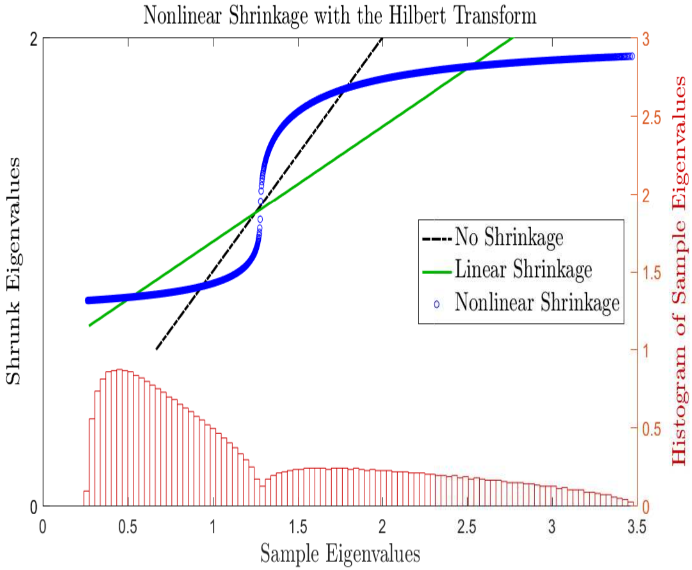
This phenomenon is easier to see based on the 'scaled' density $\varphi(x) := \pi x f(x)$, which yields the equivalent oracle formula

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{o}(x) = \frac{x}{1 + c^2\left[\varphi^2(x) + H_{\varphi}^2(x)\right] - 2cH_{\varphi}(x)}$$

Crucial advantage over global, linear shrinkage:

- Sample eigenvalues may be moved away from the grand mean, towards a local mass center
- It is helpful to consider an illustration:
  H is a two-point mass at $\{0.8, 2.0\}$, $n = 18,000$, and $p = 4,000$

Illustration



From Oracle to Feasible

The oracle estimator $\widehat{\Sigma}_n^{\text{o}} := U_n \Delta^{\text{o}} U_n'$ is not available in practice.

A bona fide estimator that also minimizes the asymptotic loss could be obtained via uniformly consistent estimation of δ^{o}.

Previous Approaches

QuEST:

- Indirect estimation of δ^o
- Proposed by Ledoit and Wolf (2012, AoS; 2015, JMVA)
- First find consistent estimator $\widehat{H}_n$ of H
- Then feed $\widehat{H}_n$ into the Marčenko-Pastur equation, together with $\widehat{c}_n := p/n$, and make us of the resulting $\widehat{F}_n$
- Difficult to implement and slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

Previous Approaches

QuEST:

- Indirect estimation of δ^{o}
- Proposed by Ledoit and Wolf (2012, AoS; 2015, JMVA)
- First find consistent estimator $\widehat{H}_n$ of H
- Then feed $\widehat{H}_n$ into the Marčenko-Pastur equation, together with $\widehat{c}_n := p/n$, and make us of the resulting $\widehat{F}_n$
- Difficult to implement and slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

NERCOME:

- Proposed by Abadir et al. (2104, JoE) and Lam (2106, AoS)
- Based on repeated sample splits to estimate the two components of $\delta^{*}_{n,i} := u'_{n,i}\Sigma_n u_{n,i}$ separately
- Requires brute-force spectral decomposition of many matrices
- Easy to implement but also slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

New Approach: Direct Estimation

Recall:

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\mathrm{o}}(x) := \frac{x}{\left[\pi c x f(x)\right]^2 + \left[1 - c - \pi c x \mathcal{H}_f(x)\right]^2}$$

Therefore, uniformly consistent estimation of δ^{o} can be based on:

(i) consistent estimation of c

(ii) uniformly consistent estimation of f

(iii) uniformly consistent estimation of $\mathcal{H}_f$

New Approach: Direct Estimation

Recall:

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{o}(x) := \frac{x}{\left[\pi c x f(x)\right]^2 + \left[1 - c - \pi c x \mathcal{H}_f(x)\right]^2}$$

Therefore, uniformly consistent estimation of δ^{o} can be based on:

(i) consistent estimation of c

(ii) uniformly consistent estimation of f

(iii) uniformly consistent estimation of $\mathcal{H}_f$

Problem (i) is trivially solved by using $\widehat{c_n} := p/n$.

Problems (ii) and (iii) can be solved by kernel estimation.

New Approach: Direct Estimation

Recall:

$$\forall x \in \mathsf{Supp}(F) \qquad \delta^{\mathrm{o}}(x) := \frac{x}{\left[\pi c x f(x)\right]^2 + \left[1 - c - \pi c x \mathcal{H}_f(x)\right]^2}$$

Therefore, uniformly consistent estimation of δ^{o} can be based on:

(i) consistent estimation of c

(ii) uniformly consistent estimation of f

(iii) uniformly consistent estimation of $\mathcal{H}_f$

Problem (i) is trivially solved by using $\widehat{c_n} := p/n$.

Problems (ii) and (iii) can be solved by kernel estimation.

Advantages:

- Easy to implement and fast to execute
- Can go to at least dimension $p = 10,000$ computationally

Outline

Choice of Kernel

A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- k is a continuous, symmetric density with finite support, mean zero, and variance one
- Its Hilbert transform $\mathcal{H}_k$ exists and is continuous
- Both the kernel k and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation

Choice of Kernel

A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- k is a continuous, symmetric density with finite support, mean zero, and variance one
- Its Hilbert transform $\mathcal{H}_k$ exists and is continuous
- Both the kernel k and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation

We use the semi-circle kernel dating back to Wigner (1955, AoM).

Choice of Kernel

A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- k is a continuous, symmetric density with finite support, mean zero, and variance one
- Its Hilbert transform $\mathcal{H}_k$ exists and is continuous
- Both the kernel k and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation

We use the semi-circle kernel dating back to Wigner (1955, AoM).

Of the 48 elementary functions whose Hilbert transform is known in closed form, it is the only one satisfying all the above assumptions.

Choice of Bandwidth

We propose to use a variable bandwidth that is proportional to the magnitude of a given sample eigenvalue.

The bandwidth applied to $\lambda_{n,i}$ is $h_{n,i} := \lambda_{n,i} h_n$, where $h_n \to 0$.

We use $h_n := n^{-0.35}$, close to the choice $n^{-1/3}$ by Jing et al. (2010, AoS). (Although they use a uniform bandwidth $h_{n,i} \equiv n^{-1/3}$).

Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of f and $\mathcal{H}_f$:

$$\forall x \in \mathbb{R} \qquad \widetilde{f}_n(x) := \frac{1}{p}\sum_{i=1}^{p} \frac{1}{h_{n,i}} k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right)$$

Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of f and $\mathcal{H}_f$:

$$\forall x \in \mathbb{R} \qquad \widetilde{f}_n(x) := \frac{1}{p}\sum_{i=1}^{p}\frac{1}{h_{n,i}}k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right)$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widetilde{f}_n}(x) := \frac{1}{p}\sum_{i=1}^{p}\frac{1}{h_{n,i}}\mathcal{H}_k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{\pi}PV\int\frac{\widetilde{f}_n(t)}{x-t}dt$$

Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of f and $\mathcal{H}_f$:

$$\forall x \in \mathbb{R} \qquad \widetilde{f}_n(x) := \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} k\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right)$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widetilde{f}_n}(x) := \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} \mathcal{H}_k\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{\pi} PV \int \frac{\widetilde{f}_n(t)}{x - t} dt$$

Feasible nonlinear shrinkage estimation:

$$\forall x \in \mathsf{Supp}(F) \qquad \widetilde{\delta}_n(x) := \frac{x}{\left[\pi \widehat{c}_n x \widetilde{f}_n(x)\right]^2 + \left[1 - \widehat{c}_n - \pi \widehat{c}_n x \mathcal{H}_{\widetilde{f}_n}(x)\right]^2}$$

Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of f and $\mathcal{H}_f$:

$$\forall x \in \mathbb{R} \qquad \widetilde{f}_n(x) := \frac{1}{p}\sum_{i=1}^{p} \frac{1}{h_{n,i}} k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right)$$

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widetilde{f}_n}(x) := \frac{1}{p}\sum_{i=1}^{p} \frac{1}{h_{n,i}} \mathcal{H}_k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{\pi} PV \int \frac{\widetilde{f}_n(t)}{x-t} dt$$

Feasible nonlinear shrinkage estimation:

$$\forall x \in \mathsf{Supp}(F) \qquad \widetilde{\delta}_n(x) := \frac{x}{\left[\pi \widehat{c}_n x \widetilde{f}_n(x)\right]^2 + \left[1 - \widehat{c}_n - \pi \widehat{c}_n x \mathcal{H}_{\widetilde{f}_n}(x)\right]^2}$$

$$\widetilde{\Sigma}_n := U_n \widetilde{\Delta}_n U_n'$$

Outline

Executive Summary

Performance of direct nonlinear shrinkage:

- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME

Executive Summary

Performance of direct nonlinear shrinkage:

- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME

Speed of direct nonlinear shrinkage:

- Basically as fast as linear shrinkage
- Much faster than QuEST
- Much faster than NERCOME

Executive Summary

Performance of direct nonlinear shrinkage:

- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME

Speed of direct nonlinear shrinkage:

- Basically as fast as linear shrinkage
- Much faster than QuEST
- Much faster than NERCOME

$\Longrightarrow$ Get the best of both worlds!

Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

$$\mathrm{PRIAL}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n\right) := \frac{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right)\right]}{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n^*, \Sigma_n\right)\right]} \times 100\%$$

Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

$$\mathrm{PRIAL}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n\right) := \frac{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right)\right]}{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n^*, \Sigma_n\right)\right]} \times 100\%$$

By construction:

- Sample covariance matrix has $\mathrm{PRIAL}_n^{\mathrm{MV}}\left(S_n\right) = 0\%$
- Finite-sample optimal estimator has $\mathrm{PRIAL}_n^{\mathrm{MV}}\left(S_n^*\right) = 100\%$

Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

$$\mathrm{PRIAL}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n\right) := \frac{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(\widehat{\Sigma}_n, \Sigma_n\right)\right]}{\mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^{\mathrm{MV}}\left(S_n^*, \Sigma_n\right)\right]} \times 100\%$$

By construction:

- Sample covariance matrix has $\mathrm{PRIAL}_n^{\mathrm{MV}}\left(S_n\right) = 0\%$
- Finite-sample optimal estimator has $\mathrm{PRIAL}_n^{\mathrm{MV}}\left(S_n^*\right) = 100\%$

Note:

- Negative PRIAL values are possible

Baseline Scenario

We use a scenario introduced by Bai and Silverstein (1998, AoP):

- Dimension $p = 200$
- Sample size $n = 600$
- Concentration ratio $\widehat{c_n} = 1/3$
- 20% of the $\tau_{n,i}$ are equal to 1, 40% equal to 3, and 40% equal to 10
- Condition number $\theta = 10$
- Variates are normally distributed

Baseline Scenario

We use a scenario introduced by Bai and Silverstein (1998, AoP):

- Dimension $p = 200$
- Sample size $n = 600$
- Concentration ratio $\widehat{c_n} = 1/3$
- 20% of the $\tau_{n,i}$ are equal to 1, 40% equal to 3, and 40% equal to 10
- Condition number $\theta = 10$
- Variates are normally distributed

Each feature will be varied in subsequent scenarios.

Results for Baseline Scenario

Estimator	Sample	Linear	Direct	QuEST	NERCOME	FSOPT
∅ Loss	2.71	2.10	1.52	1.50	1.58	1.48
PRIAL	0%	50%	97%	98%	92%	100%
Time (ms)	1	3	4	2,233	2,990	3

Note:

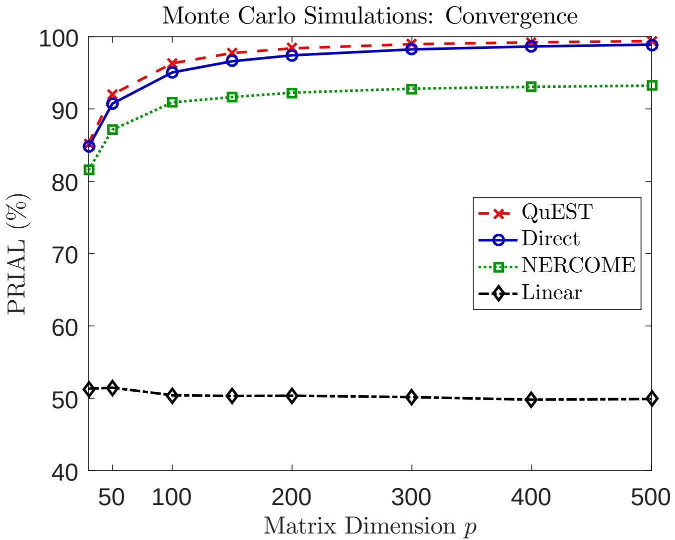
- Computational times in milliseconds come from a 64-bit, quad-core 4.00GHz Windows PC running Matlab R2016a

Large-Dimensional Asymptotics

Let p and n go to infinity together with $p/n \equiv 1/3$:

Large-Dimensional Asymptotics

Let p and n go to infinity together with $p/n \equiv 1/3$:

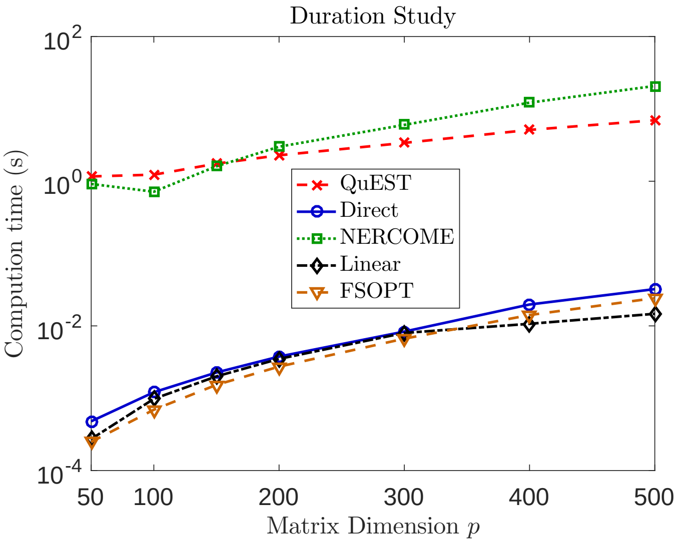


Speed

Let p and n go to infinity together with $p/n \equiv 1/3$:

Speed

Let p and n go to infinity together with $p/n \equiv 1/3$:



Ultra-High Dimension

Repeat baseline scenario but multiply both p and n by 50:

- $p = 10,000$
- $n = 30,000$

QuEST and NERCOME are no longer computationally feasible.

Ultra-High Dimension

Repeat baseline scenario but multiply both p and n by 50:

- $p = 10,000$
- $n = 30,000$

QuEST and NERCOME are no longer computationally feasible.

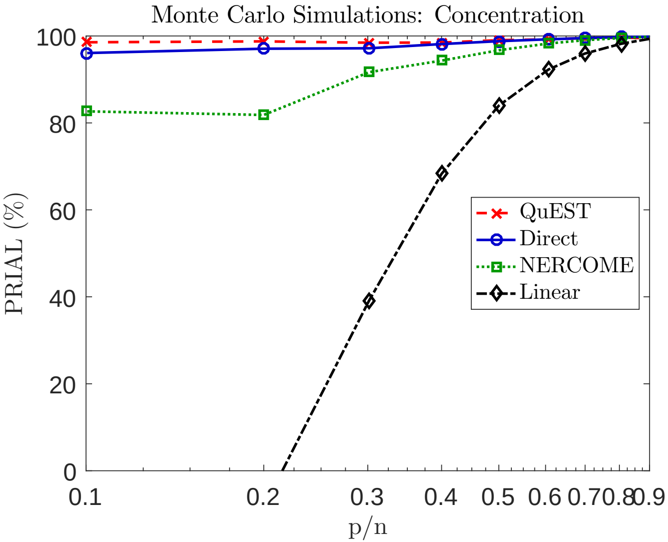
Estimator	Sample	Linear	Direct	FSOPT
∅ Loss	2.679	2.086	1.488	1.487
PRIAL	0%	49.74%	99.92%	100%
Time (s)	21	43	113	108

Concentration Ratio

Vary p/n from 0.1 to 0.9 while keeping $p \times n = 120,000$:

Concentration Ratio

Vary p/n from 0.1 to 0.9 while keeping $p \times n = 120,000$:

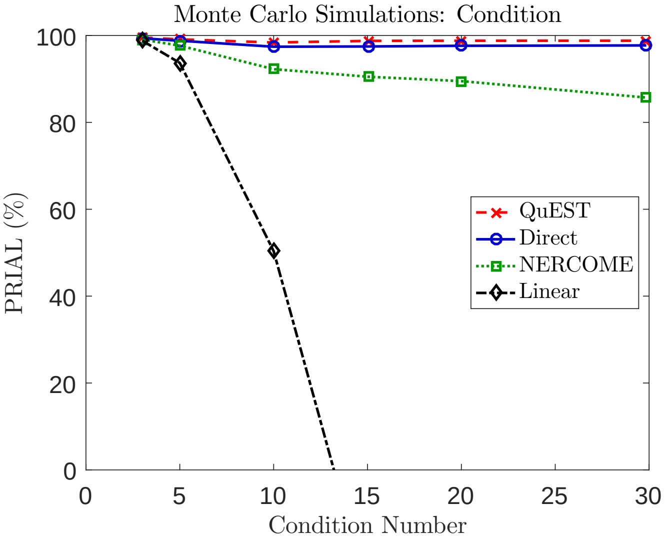


Condition Number

Vary θ from 3 to 30, by linearly squeezing/stretching the $\tau_{n,i}$:

Condition Number

Vary θ from 3 to 30, by linearly squeezing/stretching the $\tau_{n,i}$:



Non-Normality

Vary the distribution of the variates:

Non-Normality

Vary the distribution of the variates:

Distribution	Linear	Direct	QuEST	NERCOME
Normal	50%	97%	98%	92%
Bernoulli	51%	98%	98%	92%
Laplace	50%	97%	98%	92%
‘Student’ t_5	49%	97%	98%	92%

Shape of the Distribution of Population Eigenvalues

Use a shifted and stretched Beta distribution with support [1,10]:

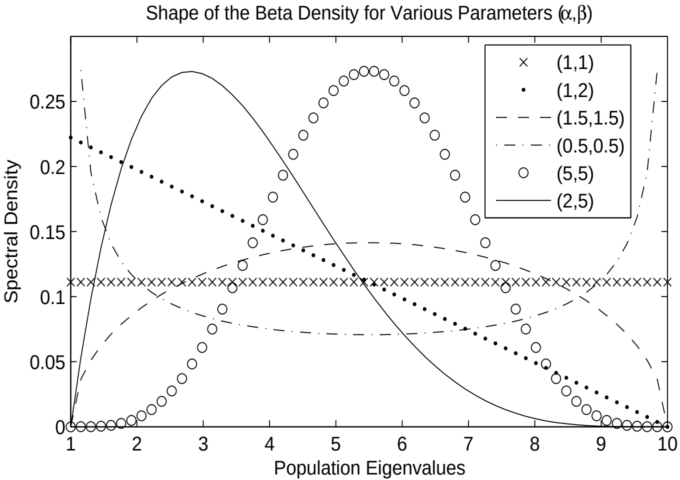
Shape of the Distribution of Population Eigenvalues

Use a shifted and stretched Beta distribution with support [1,10]:

Beta Parameters	Linear	Direct	QuEST	NERCOME
$(1, 1)$	83%	98%	99%	96%
$(1, 2)$	95%	99%	99%	98%
$(2, 1)$	94%	99%	99%	99%
$(1.5, 1.5)$	92%	99%	99%	98%
$(0.5, 0.5)$	50%	98%	98%	94%
$(5, 5)$	98%	100%	100%	99%
$(5, 2)$	97%	100%	100%	98%
$(2, 5)$	99%	99%	99%	99%

Shape

Selected (shifted and stretched) beta densities used:

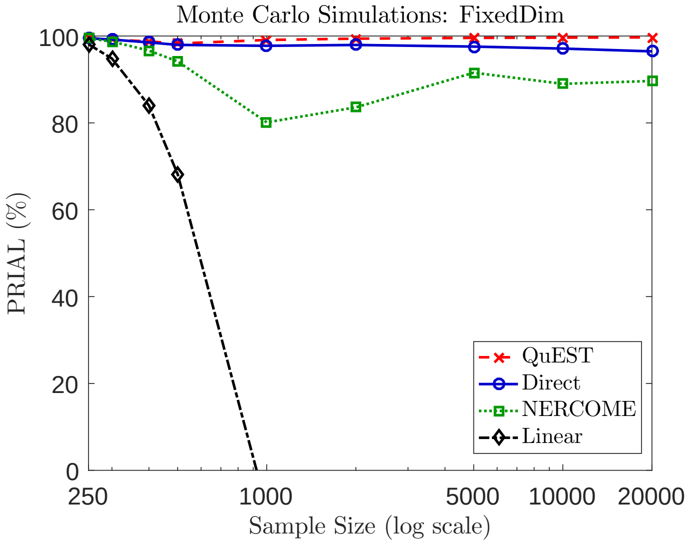


Fixed-Dimensional Asymptotics

Let n grow from 250 to 20,000 while keeping $p \equiv 200$:

Fixed-Dimensional Asymptotics

Let n grow from 250 to 20,000 while keeping $p \equiv 200$:

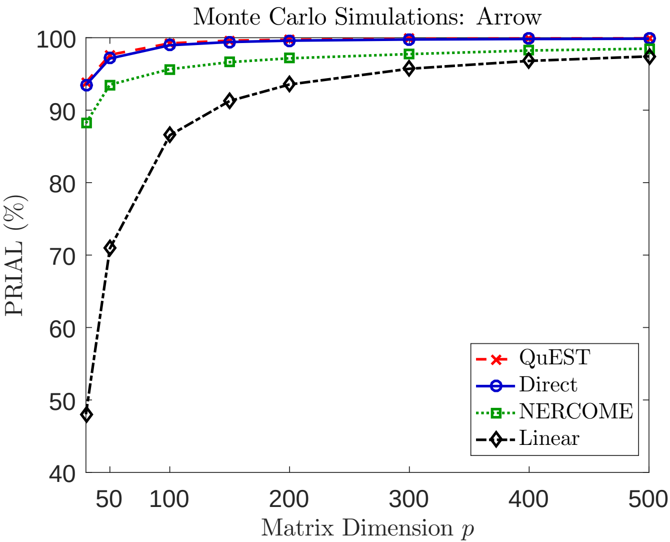


Arrow Model

Let $\tau_{n,p} := 1 + 0.5(p - 1)$ and remaining bulk from s&s Beta(5,2):

Arrow Model

Let $\tau_{n,p} := 1 + 0.5(p-1)$ and remaining bulk from s&s Beta(5,2):



Outline

Conclusion

Nonlinear shrinkage estimation of covariance matrices is a complex, but powerful structure-free approach in large dimensions.

Existing methods are difficult to implement, computationally expensive, or even both.

We have suggested a direct method based on kernel estimation that (i) performs as well as existing methods and (ii) is computationally as cheap as linear shrinkage.

Conclusion

Nonlinear shrinkage estimation of covariance matrices is a complex, but powerful structure-free approach in large dimensions.

Existing methods are difficult to implement, computationally expensive, or even both.

We have suggested a direct method based on kernel estimation that (i) performs as well as existing methods and (ii) is computationally as cheap as linear shrinkage.

This direct method also can handle dimensions of +1 magnitude, which is a big + in the age of Big Data.

Abadir, K., Distaso, W., and Žikesš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181:165–180.

Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the suppport of the limiting spectral distribution of large-dimensional random matrices. *Annals of Probability*, 26(1):316–345.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B*, 75(4):603–680.

Jing, B.-Y., Pan, G., Shao, Q.-M., and Zhou, W. (2010). Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Annals of Statistics*, 38(6):3724–3750.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics*, 44(3):928–953.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139(2):360–384.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Annals of Statistics*, 36(6):2818–2849.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564.